

SberQuAD – Russian Reading Comprehension Dataset: Description and Analysis

Pavel Efimov^{1*}, Andrey Chertok², Leonid Boytsov, Pavel Braslavski^{3,4}

¹Saint Petersburg State University, Saint Petersburg, Russia

²Sberbank, Moscow, Russia

³Ural Federal University, Yekaterinburg, Russia

⁴JetBrains Research, Saint Petersburg, Russia

pavel.vl.efimov@gmail.com, achertok@sberbank.ru, leo@boytsov.info, pbras@yandex.ru

Abstract

SberQuAD—a large scale analog of Stanford SQuAD in the Russian language—is a valuable resource that has not been properly presented to the scientific community. We fill this gap by providing a description, a thorough analysis, and baseline experimental results.

Keywords: reading comprehension, question answering, Russian language resources, evaluation

1. Introduction

On September 14, 2017 a data science department of Sberbank¹—the largest financial institution in Russia—announced a question answering (QA) challenge with substantial monetary prizes. For this competition Sberbank provided a new large Russian QA dataset containing about 50K training examples, 15K development, and 25K testing examples (see § 3. for a detailed description). It was created similarly to the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016), which is reflected in its name *SberQuAD* (Sberbank Question Answering Dataset). The competitions had two tasks: retrieval of answer-bearing paragraphs and a reading comprehension (RC) task, which is the focus of this study.

Despite high participation—during a 1.5-month competition 120 participants made 1,348 submissions—the dataset and the contest were neither properly documented nor presented to the scientific community: We were able to find only two studies using SberQuAD (Kuratov and Arkhipov, 2019; Soboleva and Vorontsov, 2019). Given the importance of the RC task, the scarcity of non-English resources, and the amount of effort went into creation of SberQuAD, it is important to fill the gap. We in turn provide a historical overview, a description, an in-depth analysis, and baseline experimental results for SberQuAD (using methods previously applied to SQuAD). We believe this is an important contribution to research in multilingual QA.

2. Related Work

QA tasks for unstructured data are typically divided into open-domain (Prager, 2006) and story comprehension tasks (Hirschman and Gaizauskas, 2001). In the open-domain setting, to answer a question the system first needs to guess which documents may contain answers. The modern history of open-domain QA starts from TREC challenges organized by NIST in 2000s (Dang et al., 2007) and extended by CLEF to a multilingual setting (Giampiccolo et al., 2008). Notably,

in 2011 the open-domain system IBM Watson outstripped two human champions in the QA contest Jeopardy! (Ferrucci et al., 2010).

The story comprehension—commonly referred to as reading comprehension (RC)—is a more restricted task, where the system needs to answer questions for a given document. This task has recently become quite popular with the introduction of a large scale RC dataset named SQuAD (Rajpurkar et al., 2016), which was created by crowd workers. The dataset contains more than 100K questions posed to paragraphs from popular Wikipedia articles. An answer to each question should be a valid and relevant paragraph phrase, i.e., a contiguous sequence of paragraph words including but not limited to named entities and noun phrases. The second version of SQuAD (SQuAD 2.0) contains a number of unanswerable questions (Rajpurkar et al., 2018). This makes the task more difficult as the system needs to figure out when an answer does not exist.

Wide adoption of SQuAD led to emergence of many datasets. TriviaQA (Joshi et al., 2017) consists of 96K trivia game questions and answers found online accompanied by answer-bearing documents. Natural Questions dataset (Kwiatkowski et al., 2019) is approximately three times larger than SQuAD. In that, unlike SQuAD, questions are sampled from Google search log rather than generated by crowd workers. MS MARCO (Bajaj et al., 2016) contains 1M questions from a Bing search log along with free-form answers. For both MS MARCO and Natural Questions answers are produced by in-house annotators. QuAC (Choi et al., 2018) and CoQA (Reddy et al., 2019) contain questions and answers in information-seeking dialogues. For a more detailed discussion we address the reader to a recent survey (Zhang et al., 2019).

Majority of RC dataset are in English. Few exceptions are Chinese datasets WebQA (Li et al., 2016) and DuReader (He et al., 2017), as well as Bulgarian (Hardalov et al., 2019) and Tibetan (Sun and Xia, 2019) ones. Recently, Artetxe et al. (2019) experimented with cross-language transfer learning and prepared XQuAD dataset containing 240 paragraphs and 1,190 question-answer pairs from SQuAD v1.1 translated

*Work done as an intern at JetBrains Research.

¹<https://www.sberbank.com/about>

<i>P6418</i>	Термин Computer science (Компьютерная наука) появился <u>в 1959 году</u> в научном журнале Communications of the ACM, в котором <u>Луи Фейн</u> (Louis Fein) ратовал за создание Graduate School in Computer Sciences (Высшей школы в области информатики) ... Усилия Луи Фейна, численного аналитика Джорджа Форсайта и других увенчались успехом: университеты пошли на создание программ, связанных с информатикой, начиная <u>с Университета Пердью</u> в 1962.
<i>P6418</i>	The term "computer science" appears in a <u>1959</u> article in Communications of the ACM, in which <u>Louis Fein</u> argues for the creation of a Graduate School in Computer Science ... Louis Fein's efforts, and those of others such as numerical analyst George Forsythe, were rewarded: universities went on to create such departments, starting with <u>Purdue</u> in 1962.
<i>Q11870</i>	Когда впервые был применен термин Computer science (Компьютерная наука)?
<i>Q11870</i>	When did the term "computer science" appear?
<i>Q28900</i>	Кто впервые использовал этот термин?
<i>Q28900</i>	Who was the first to use this term?
<i>Q30330</i>	Начиная с <u>каого</u> учебного заведения стали применяться учебные программы, связанные с информатикой?
<i>Q30330</i>	Starting with <u>wich</u> university were computer science programs created?

Figure 1: A sample SberQuAD entry (both the original and the translation): answers are underlined and colored. The word **which** in *Q30330* is misspelled on purpose to reflect the fact that the original has a misspelling.

into 10 languages.

A number of studies scrutinize existing datasets to evaluate the difficulty of the task as well as robustness of the models. Chen et al. (2016) sample 100 passage-question-answer triples from CNN/Daily Mail dataset (Hermann et al., 2015), classify them manually according to difficulty into several categories, and compare performance of several models for different levels of question complexity. Jia and Liang (2017) generate semi-automatically distracting sentences into paragraphs to investigate the robustness of neural reading comprehension models. Talmor and Berant (2019) study how well neural network models transfer among different RC datasets. Wadhwa et al. (2018) perform quantitative and qualitative analysis of four neural models on SQuAD. This work is close to ours, though our focus is on exploring the SberQuAD dataset, rather than models.

Sugawara et al. (2017) evaluate dataset difficulty from a human perspective. They compute readability scores and the number of skills, such as reasoning and co-reference capability, required to answer questions for six datasets. They show that SQuAD is an easy-to-answer but hard-to-read dataset, which requires few skills to answer questions. Rondeau and Hazen (2018) argue that human explanations are not reliable and instead estimate question complexity based on performance of several models: The fewer models can answer a question, the higher is its complexity. They find that complexity correlates with several features including the presence of named entities and the density of question words in the answer, but not with readability scores. In particular, if the answer is an named entity the question is easy in 72% of the cases as opposed to 44% of the cases when the answer is not an entity.

3. Dataset

In this paper, we focus on task B of the Sberbank 2017 competition and the corresponding RC dataset, namely, SberQuAD. Details of the competition can be found on the competition website². The original dataset comes in

CSV format: A variant in SQuAD JSON format can be downloaded from the DeepPavlov QA project page.³ A lively discussion of approaches and their effectiveness can be found in the OpenDataScience community slack⁴. The list of top-10 best performing teams can be found in the video of the celebration ceremony.⁵

As we learned from a private communication with the dataset developers, they generally followed the procedure described in the SQuAD paper (Rajpurkar et al., 2016). First they selected Wikipedia pages, split them into paragraphs, and presented paragraphs to crowd workers. For each paragraph, a crowd worker had to come up with questions that can be answered using solely the content of the paragraph. In that, an answer must have been a paragraph span, i.e., a contiguous sequence of paragraph words. The tasks were posted on Toloka⁶ crowdsourcing platform. SberQuAD contains 50,364 paragraph-question-answer triples in the training set which are publicly available. A development set, which was available only during the competition, contains about 15K triples and the hold-out test containing 25K triples is kept private. There are two differences between SQuAD and SberQuAD formats. First, SberQuAD does not tell us which Wikipedia pages a paragraph belongs to. Second, each answer is represented by a string without respective starting position in the paragraph.

Examples and basic statistics. Figure 1 shows a sample SberQuAD paragraph with three questions: Gold-truth answers are underlined in text. Generally, the format of the question and the answers mimics that of SQuAD v1.1. Note, however, the following peculiarities: Question *Q30330* contains a spelling error; Question *Q28900* references prior question *Q11870* and cannot, thus, be answered on its

²<https://github.com/sberbank-ai/data-science-journey-2017>

³<http://docs.deeppavlov.ai/en/master/features/models/squad.html> DeepPavlov is an open NLP framework maintained by the Neural Networks and Deep Learning Lab of Moscow Institute of Physics and Technology (MIPT), see (Burtsev et al., 2018).

⁴<https://opendatascience.slack.com/>, hashtag #sberbank_contest, September–November 2017.

⁵https://youtu.be/J5H0jC4Xn_Y?t=29830

⁶<https://toloka.yandex.com>

	<i>SberQuAD</i> train	<i>SQuAD 1.1</i> train/dev
# questions	50,364	87,599 / 10,570
# unique paragraphs	9,080	18,896 / 2,067
Number of tokens		
avg. paragraph length	101.7	116.6 / 122.8
avg. question length	8.7	10.1 / 10.2
avg. answer length	3.7	3.16 / 2.9
avg. answer position	40.5	50.9 / 52.9
Number of characters		
avg. paragraph length	753.9	735.8 / 774.3
avg. question length	64.4	59.6 / 60.0
avg. answer length	25.9	20.2 / 18.7
avg. answer position	305.2	319.9 / 330.5
question-paragraph LCMS	32.7	19.5 / 19.8

Table 1: SberQuAD statistics in the # of characters and tokens. LCMS stands for the longest contiguous matching subsequence.

own (likely both questions were created by the same crowd worker).

Basic dataset statistics is summarized in Table 1: SberQuAD has about twice as fewer questions compared to SQuAD whereas the average lengths of paragraphs, questions, and answers are very similar for both datasets. Distribution of question/answer length is presented in Figure 2. There are 275 questions (0.55%) having at least 200 characters and 374 answers (0.74%) that are longer than 100 characters. Questions are substantially longer than answers. Anecdotaly, very long answers and very short questions are frequently errors. For example, for question *Q61603* the answer field contains a copy of the whole paragraph, while question *Q76754* consists of a single word ‘thermodynamics’.

Because the original SberQuAD development set is not available, the original training set of SberQuAD was partitioned into a (new) training (45,328) and testing (5,036) sets by the DeepPavlov team. This is the partition that we use in our experiments: We train models on the training set and evaluate them on the testing set.

Analysis of questions. Most questions in the dataset start with either a question word or preposition: ten most common starting words are что (what), в (in), как (how), кто (who), какие (what adj), когда (when), какой (what adj), где (where), сколько (how many), на (on).

These starting words correspond to 62.4% of all questions. In about 4% of the cases, an interrogative word is not among the first three words of the question, though. Manual inspection showed that in most cases these entries are declarative statements, sometimes followed by a question mark, e.g. *Q15968* ‘famous Belgian poets?’, or ungrammatical questions.

To get a better understanding of question types, we inspected questions’ most common lemmatized starting bigrams (Table 11) and trigrams (Table 12). In Russian, an interrogative word is often preceded by a preposition, which results in a high variability of starting n-grams: As one can see from the Table 11, ten most frequent bigrams account only for about

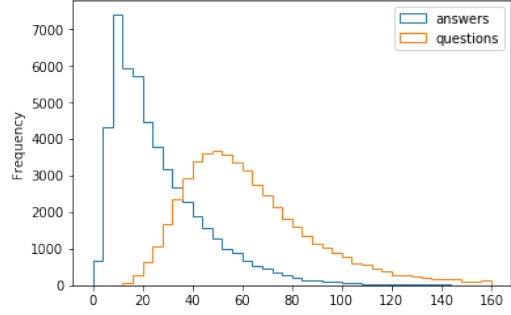


Figure 2: Question/answer length histograms (in chars)

19% of all questions. Judging by bigram statistics, definition (*what do you call/what is X*) and time-related questions (*when*) are among most popular ones. Trigram statistics (Table 12) permits a more precise inference about most common question types: They include variations of time-related questions (*in which year/century/period*), location questions (*in which city/country*), as well as causality questions (*what does X lead to/what does X depend on*).

Analysis of answers. While manually examining the dataset, we encountered misspelled questions. To estimate the proportion of questions with misspellings, we verified all questions using Yandex spellchecking API⁷. The automatic speller identified 2,646 and 287 misspelled questions in training and testing sets, respectively. According to a manual assessment of 200 randomly selected questions, the spellchecker has precision 0.62.⁸ Manual inspection suggests that most false positives are due to either spelling/inflectional variants or rare words being replaced with more frequent ones (apparently based on language model scores).

We also found 385 and 51 questions in training and testing sets, respectively, containing Russian interrogative particle *ли* (*whether/if*). This form implies a yes/no question, which is generally not possible to answer in the RC setting by selecting a valid and relevant paragraph phrase. For this reason, most answers for these yes-no questions are fragments supporting or refuting the question statement. In addition, we found 15 answers in the training set, where the correct answer ‘yes’ (Russian *да*) can be found as a paragraph word substring, but not as a valid/relevant phrase.

Following (Rajpurkar et al., 2016), we analyzed answers presented in the dataset by their type. To this end, we employed a NER tool from DeepPavlov library, DPNER hereafter.⁹ DPNER is a multilingual BERT model trained on OntoNotes corpus annotated with 19 entity types and transferred to Russian (for a discussion of zero-shot transfer see a paper by Pires et al. (2019)). To evaluate DPNER on SberQuAD data, we randomly sampled 1,000 answers and manually tagged containing named entities (NE) using the

⁷<https://yandex.ru/dev/speller/> (in Russian)

⁸This score is significantly lower than scores obtained for Yandex spellchecker by DeepPavlov ($P = 83.09$; $R = 59.86$), see http://docs.deeppavlov.ai/en/master/features/models/spelling_correction.html

⁹<http://docs.deeppavlov.ai/en/master/features/models/ner.html>

NE	Manual	DPNER (P/R)	Exact
Date	9.9%	12.6% (0.83/0.96)	2.31%
Number	11.0%	9.9% (0.85/0.90)	3.37%
Person	8.8%	8.2% (0.89/0.89)	3.85%
Location	5.4%	7.6% (0.64/0.87)	1.45%
Organization	4.0%	3.6% (0.70/0.70)	1.46%
Other NEs	5.2%	2.5% (0.71/0.44)	0.97%

Table 2: Named entities in a manually annotated sample of 1,000 answers (**Manual**); answers containing automatically detected NEs (**DPNER**); detection quality on manually annotated sample (**P/R**); automatically detected NEs that exactly match answers’ boundaries (**Exact**).

Type	% test	R-Net	BiDAF	DocQA	DrQA	BERT
NP	24.0	77.5	70.3	78.2	73.5	84.5
PP	10.5	83.1	78.6	84.9	81.4	89.1
VP	7.1	61.9	54.0	62.7	55.5	71.6
ADJP	5.9	73.0	65.3	75.5	67.2	80.5
ADVP	0.3	67.9	45.3	70.7	51.2	76.6
non-R	0.3	91.7	88.2	98.2	92.9	95.1
None	9.1	75.7	69.0	77.1	70.1	83.0
Test set		77.8	72.2	79.5	75.0	84.8

Table 3: Distribution of answers by constituent types (NP – noun phrase, PP – prepositional phrase, VP – verb phrase, ADJP – adjective phrase, ADVP – adverb phrase, non-R – words in non-Russian characters; None – not recognized).

following tags: DATE, NUMBER, PERSON, LOCATION, ORGANIZATION, and OTHER (artwork, TV show, etc.). When an answer contained more than one entity, we highlighted the entity representing a key answer concept (e.g., the head noun phrase). For example *26-year-old Comtesse Sophie d’Houdetot* (Q56395) was marked as PERSON.

This statistics is summarized in Table 2. The first column shows distribution of NEs in answers according to manual annotation. The second column shows precision/recall of the NER tool applied to sentences containing answers. The last column of the Table 2 reports a fraction of answers (identified by DPNER) that are NE (as opposed only a part of the answer being a NE). In total, DPNER found NEs in almost 43% of answers in the dataset.

Following (Rajpurkar et al., 2016), we complemented our analysis of answers with syntactic parsing. To this end we applied the rule-based constituency parser AOT¹⁰ to answers without detected NE. When AOT produced multiple parses, we picked the parse with the longest span. AOT parser supports a long list of phrase types (57 in total), we grouped them into conventional high-level types, which are shown in Table 3¹¹. Not surprisingly, noun phrases are most frequent answer types, followed by prepositional phrases. Verb phrases represent a non-negligible share of answers (7.1%), which is quite different from a traditional QA setting where answers are predominantly noun phrases (Prager, 2006).

¹⁰<http://aot.ru>

¹¹Table 3 provides data for the testing set, but the distribution for the training set is quite similar.

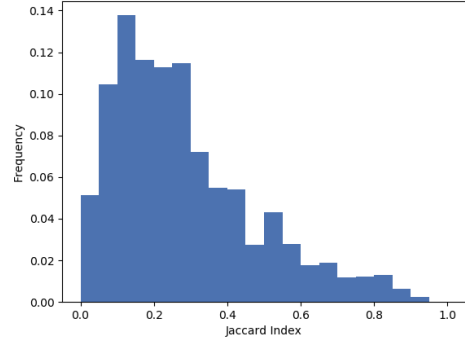


Figure 3: Jaccard similarity distribution between questions and answer containing sentences.

Question/paragraph similarity. We further estimate similarity between questions and paragraph sentences containing the answer: The more similar is the question to its answer’s context, the simpler is the task of locating the answer. In contrast to SQuAD we refrain from syntactic parsing and rely on simpler approaches. First, we compared questions with complete paragraphs. To this end, we calculated the length of the the longest contiguous matching subsequence (LCMS) between a question and a paragraph using the `diffliib` library.¹² The last row in Table 1 shows that despite similar paragraph and question lengths in both SQuAD and SberQuAD, the SberQuAD questions are more similar to the paragraph text.

Second, we estimated similarity between a question and the sentence containing the answer. To ensure the accuracy of estimation, we evaluated several available tools for sentence splitting on a random sample of 100 SberQuAD paragraphs, which were manually split into 590 sentences. DeepPavlov tokenizer¹³ outperformed other tools in terms of quality ($P/R = 0.93/0.94$) and efficiency, so we applied it to the whole train subset. Subsequently, we lemmatized the data using *mystem*¹⁴ and calculated the Jaccard coefficient between a question and the sentence containing the answer. The distribution of the scores is presented in Figure 3. The mean value of the Jaccard coefficient is 0.28 (median is 0.23).

Our analysis shows that there is a substantial lexical overlap between questions and paragraph sentences containing the answer, which may indicate a heavier use of the copy-and-paste approach by crowd workers recruited for SberQuAD creation.¹⁵

¹²<https://docs.python.org/3/library/difflib.html>

¹³https://github.com/deepmpt/ru_sentence_tokenizer

¹⁴<https://yandex.ru/dev/mystem/> (in Russian)

¹⁵Note that in the interface for crowdsourcing SQuAD questions, prompts at each screen reminded the workers to formulate questions in their own words; in addition, the copy-paste functionality for the paragraph was purposefully disabled.

NE	% test	R-Net	BiDAF	DocQA	DrQA	BERT
Date	12.2%	88.0	86.6	90.0	88.9	91.3
Number	9.6%	73.1	69.1	75.5	72.5	80.4
Person	8.8%	78.3	73.1	81.0	77.7	86.6
Location	7.6%	79.8	75.7	81.1	77.8	85.8
Organization	4.1%	79.0	77.3	82.3	78.3	88.2
Other NE	2.1%	72.7	59.4	73.6	64.7	80.9
Any NE	42.7%	80.3	76.4	82.6	79.7	87.0
Test set		77.8	72.2	79.5	75.0	84.8

Table 4: Model performance on answers containing named entities.

4. Employed Models

We used the following models:

- Two baselines provided by SberQuAD organizers;
- Four models, which have strong performance among models not relying on transformers. They were used in a study similar to ours (Wadhwa et al., 2018);
- BERT model provided by the DeepPavlov library, which employs large pre-trained transformers (Devlin et al., 2018; Vaswani et al., 2017).

Preprocessing and training. We tokenized text using spaCy¹⁶. To initialize the embedding layer for BiDAF, DocQA, DrQA, and R-Net we use Russian case-sensitive fastText embeddings trained on Common Crawl and Wikipedia¹⁷. This initialization is used for both questions and paragraphs. For BiDAF and DocQA about 10% of answer strings in both training and testing sets require a correction of positions, which can be nearly always achieved automatically by ignoring punctuation (12 answers required a manual intervention). Models were trained on GPU nVidia Tesla V100 16Gb. We used default implementation settings, which are listed in Table 5:

Model	Optim.	Batch	# epochs	Init. LR
R-Net	Adadelta	32	40 (60K steps)	0.5
BiDAF	Adadelta	60	12	0.5
DocQA	Adadelta	45	26	1
DrQA	Adamax	128	40	N/A

Table 5: Training parameters. LR stands for learning rate.

Baselines. Contest organizers made two baselines¹⁸ available. **Simple baseline:** The model returns a sentence with the maximum word overlap with the question. **ML baseline** generates features for all word spans in the sentence returned by the **simple baseline**. The feature set includes TF-IDF scores, span length, distance to the beginning/end of the sentence, as well as POS tags. The model uses gradient boosting to predict F1 score. At the testing stage the model selects a candidate span with maximum predicted score.

¹⁶<https://github.com/buriy/spacy-ru>

¹⁷<https://fasttext.cc/docs/en/crawl-vectors.html>

¹⁸https://github.com/sberbank-ai/data-science-journey-2017/tree/master/problem_B/

NE	% test	R-Net	BiDAF	DocQA	DrQA	BERT
Date	2.2%	87.1	87.3	90.8	87.5	95.0
Number	3.3%	78.2	72.4	80.1	77.7	90.2
Person	4.2%	83.2	74.0	85.1	82.9	91.4
Location	1.7%	78.3	72.8	82.1	77.9	88.6
Organization	1.5%	80.7	76.5	81.6	79.2	91.8
Other NE	0.9%	80.9	54.9	78.1	66.4	88.9
Any NE	13.8%	81.6	74.5	83.6	80.2	91.2
Test set		77.8	72.2	79.5	75.0	84.8

Table 6: Model performance on answers matching NER tags.

Gated Self-Matching Networks (R-Net): This model, proposed by Wang et al. (2017), is a multi-layer end-to-end neural network that uses a gated attention mechanism to give different levels of importance to different paragraph parts. It also uses self-matching attention for the context to aggregate evidence from the entire paragraph to refine the query-aware context representation. We use a model implementation by HKUST¹⁹. To increase efficiency, the implementation adopts scaled multiplicative attention instead of additive attention and uses variational dropout.

Bi-Directional Attention Flow (BiDAF): The model proposed by Seo et al. (2016) takes inputs of different granularity (character, word and phrase) to obtain a query-aware context representation without previous summarization using memory-less context-to-query (C2Q) and query-to-context (Q2C) attention. We use original implementation by AI2²⁰.

Multi-Paragraph Reading Comprehension (DocQA): This model, proposed by Clark and Gardner (2017), aims to answer questions based on entire documents (multiple paragraphs). If considering the given paragraph as the document, it also shows good results on SQuAD. It uses the bi-directional attention mechanism from the BiDAF and a layer of residual self-attention. We also use original implementation by AI2²¹.

Document Reader (DrQA): This model proposed by Chen et al. (2017) is part of the system for answering open-domain factoid questions using Wikipedia. The Document Reader component performs well on SQuAD (skipping the document retrieval stage). The model has paragraph and question encoding layers with RNNs and an output layer. The paragraph encoding passes as input to RNN a sequence of feature vectors derived from tokens: word embedding, exact match with question word, POS/NER/TF and aligned question embedding. The implementation is developed by Facebook Research²².

Bidirectional Encoder Representations from Transformers (BERT): We use a BERT-based QA model by DeepPavlov²³. Pre-trained BERT models achieved superior performance is a variety of downstream NLP tasks, in-

¹⁹<https://github.com/HKUST-KnowComp/R-Net>

²⁰<https://github.com/allenai/bi-att-flow>

²¹<https://github.com/allenai/document-qa>

²²<https://github.com/facebookresearch/DrQA>

²³<http://docs.deeppavlov.ai/en/master/features/models/squad.html>

Model	SberQuAD		SQuAD	
	EM	F1	EM	F1
simple baseline	0.3	25.0	—	—
ML baseline	3.7	31.5	—	—
BiDAF	51.7	72.2	68.0	77.3
DrQA	54.9	75.0	70.0	79.0
R-Net	58.6	77.8	71.3	79.7
DocQA	59.6	79.5	72.1	81.1
BERT	66.6	84.8	85.1	91.8

Table 7: Model performance on SQuAD and SberQuAD; SQuAD part shows single-model scores on test set taken from respective papers.

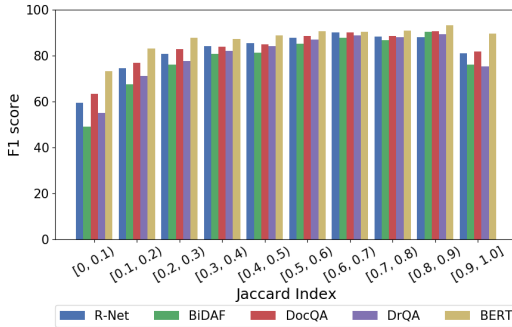


Figure 4: Model performance depending on Jaccard similarity between a question and the sentence containing an answer.

cluding RC (Devlin et al., 2018). The Russian QA model is obtained by a transfer from the multilingual BERT (mBERT) with subsequent fine-tuning on the Russian Wikipedia and SberQuAD (Kuratov and Arkhipov, 2019).

Evaluation. Similar to SQuAD, SberQuAD evaluation employs two metrics to assess model performance – 1) the percentage of system’s answers that exactly match (EM) any of the gold standard answers and 2) the maximum overlap between the system response and ground truth answer at the token level expressed via F1 (averaged over all questions). Both metrics ignore punctuation and capitalization.

5. Analysis of Model Performance

Main experimental results are shown in Table 7. It can be seen that all the models perform worse on the Russian dataset SberQuAD than on SQuAD. In that, there is a bigger difference in exact matching scores compared to F1. For example, for BERT the F1 score drops from 91.8 to 84.8 whereas the exact match score drops from 85.1 to 66.6. The relative performance of models is consistent for both datasets, although there is a greater variability among four neural “pre-BERT” models. One explanation for lower scores is that SberQuAD has always only one correct answer, whereas SQuAD can have multiple answer variants (1.7 on the development set). Furthermore, SberQuAD contains many fewer answers that are named entities than SQuAD (13.8% vs. 52.4%), which—as we discuss below—maybe another reason for lower scores. Another plausible reason is a poorer quality of annotations: We have found a number

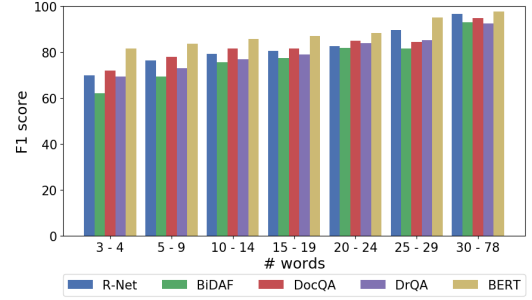


Figure 5: Model performance depending on question length (# of words).

of deficiencies including but not limited to misspellings in questions and answers.

Figure 4 shows the relationship between the F1 score and the question-answer similarity expressed as the Jaccard coefficient. Note that 64% of question-sentence pairs fall into first three bins. As expected, a higher value of the Jaccard coefficient corresponds to higher F1 scores (with the exception of 14 questions where Jaccard is above 0.9).²⁴ Furthermore, in the case of the high similarity there is only a small difference among model performance. These observations support the hypothesis that it is easier to answer questions when there is a substantial lexical overlap between a question and a paragraph sentence containing the answer.

Longer questions are easier to answer too: According to Figure 5, the F1 score increases nearly monotonically with the question length. Presumably, longer questions provide more context for identifying correct answers. In contrast, dependency on the answer length is not monotonic: the F1 score first increases and achieves the maximum for 2-4 words. A one-word ground truth constitutes a harder task: missing a single correct word results in a null F1 score, whereas returning a two-word answer containing the single correct word results in only $F1 = 0.67$. F1 score also decreases substantially for answers above average length. It can be explained by the fact that models are trained on the dataset where shorter answers prevail, see Table 1 and Figure 2. Models’ average-length answers get low scores in case of longer ground truth. For example, a 4-word answer fully overlapping with a 8-word ground truth answer gets again only $F1 = 0.67$.

Following our analysis of the dataset, we break down model scores by the answer types. Tables 4 and 6 summarize performance of the models depending on the answers containing named entities of different types. Table 4 represents answers that contain at least one NE, but which are not necessarily NEs themselves (42.7% in the test set). Table 6 represents answers that are NEs (13.8% in test). A common trend for all models is that F1 scores for answers mentioning dates, persons, locations, and organizations are higher than average. NUMBER is an exception in this regard, probably due

²⁴Among these 14 questions the majority are long sentences from the paragraph with a single word (answer) substituted by a question word; there is an exact copy with just a question mark at the end; one question has the answer erroneously attached after the very question.

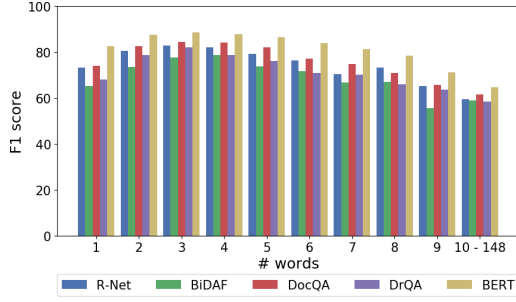


Figure 6: Model performance depending on answer length (# of words).

	% test	R-Net	BiDAF	DocQA	DrQA	BERT
w/ typos	5.7	74.1	66.7	77.5	67.5	81.1
correct	94.3	77.1	72.5	79.6	75.4	85.0
Test set		77.8	72.2	79.5	75.0	84.8

Table 8: Answer quality for misspelled questions.

to a high variability of contexts might contain numerals both as digits and words. Answers containing *other NEs* also show degraded performance – probably, again due to their higher diversity and lower counts. The scores are significantly higher when an answer is exactly a NE. This is in line with previous studies that showed that answers containing NEs are easier to answer, see for example (Rondeau and Hazen, 2018).

For about 48% of the answers in the testing set that don’t contain NEs we were able to derive their syntactic phrase type, see Table 3. Among them, non-factoid verb phrases stand out as most difficult ones (all models perform worse on such questions).²⁵ In contrast, answers expressed as prepositional phrases are easier to answer compared to both noun and verb phrases. Noun phrases—most common syntactic units among answers—are second-easiest structure among others to answer. However, with exception for BERT, F1 scores for noun phrases are lower than average.

The models behave remarkably differently on questions with and without detected misspellings, see Table 8. DrQA seems to be most sensible to misspellings: The difference in F1 is almost 8% (scores are lower for misspelled questions). DocQA has most stable behavior: The difference in F1 scores is about 2%.

Questions with interrogative *ли*-particle represent around 1% in the whole dataset. Although score averages for such small sets are not very reliable, the decrease in performance on these questions is quite sharp and consistent for all models: It ranges from 8.5% in F1 points for DocQA to 18.7% for BiDAF. We hypothesize that these questions are substantially different from other questions and are poorly represented in the training set.

Due to high variability of starting question n-grams (see Tables 11 and 12), we cannot make reliable statements for all but most frequent ones. For these—we can conclude—that model performance is mostly above average. There are a few exceptions: Notably, some variants of the definition

²⁵ Adverbial phrases appears to be even harder, but they are too few to make reliable conclusions.

	% test	R-Net	BiDAF	DocQA	DrQA	BERT
w/ <i>ли</i>	1.0	66.6	53.7	71	57.5	73.3
other	99.0	77.9	72.4	79.6	75.2	84.9
Test set		77.8	72.2	79.5	75.0	84.8

Table 9: Yes/no (*ли*-particle) questions.

Category	%
Incomplete answer	29
Vague question	19
Incorrect answer	14
Broad question	12
Co-reference resolution	12
Reasoning	10
Misspellings	6
No answer	3
Yes/no	3
Paraphrase	3

Table 10: Qualitative analysis of 100 difficult questions (questions can be assigned to more than one category).

questions *what/who is* are especially hard for BiDAF. More concrete *when*-questions appear to be an easier task for all models. In the case of trigrams the number of questions of each type is much smaller (recall that the testing set contains around 5,000 questions). Nevertheless, the scores for most frequent questions *in which year* are much better than the average scores.

Finally, we sampled 100 questions where all models achieved zero F1 score (i.e., they returned a span with no overlap with a ground truth answer). We manually grouped the sampled questions into the following categories:

- An entire paragraph or its significant part can be seen as an answer to a *broad/general question*.
- An answer is *incomplete*, because it contains only a part of an acceptable longer answer. For example for Q31929 ‘Who did notice an enemy airplane?’ only the word *pilots* is marked as ground truth in the context: *On July 15, during a reconnaissance east to Zolotaya Lipa, pilots of the 2nd Siberian Corps Air Squadron Lieutenant Pokrovsky and Cornet Plonsky noticed an enemy airplane.*
- *Vague questions* are related to the corresponding paragraph but seem to be a result of a misinterpretation of the context by a crowdsource worker. For example, in Q70465 ‘What are the disadvantages of TNT comparing to dynamite and other explosives?’ the ground truth answer ‘a detonator needs to be used’ is not mentioned as a disadvantage in the paragraph. A couple of these questions use paronyms of concepts mentioned in the paragraph. For example, Q46229 asks about ‘discrete policy’, while the paragraph mentions ‘discretionary policy’.
- *No answer in the paragraph* and *incorrect answer* constitute more straightforward error cases.

Bigram	% test	R-Net	BiDAF	DocQA	DrQA	BERT
в какой / in what	8.62	84.2	82.7	85.8	84.6	87.7
как называться / how is X called	2.46	84.5	74.7	81.9	78.7	89.8
кто быть / who was	1.21	81.8	71.0	83.2	78.3	89.2
на какой / on what	1.21	75.9	72.7	76.7	78.0	80.2
что такой / what is	1.15	71.6	67.6	74.4	70.6	77.0
с какой / with what	1.01	76.6	78.3	79.4	78.4	89.9
для что / what for	0.91	81.6	79.8	82.5	78.1	86.9
к что / to what	0.77	90.9	82.2	86.7	88.1	90.2
что являться / what is	0.69	84.6	88.0	93.5	87.2	93.2
когда быть / when was	0.68	79.2	82.2	84.0	86.9	92.5
Test set		77.8	72.2	79.5	75.0	84.8

Table 11: Model F1 scores depending on questions’ leading bigrams (bigrams are lemmatized).

Trigram	% test	R-Net	BiDAF	DocQA	DrQA	BERT
в какой год / in which year	4.39	89.4	88.8	90.4	89.9	91.0
в какой город / in which city	0.32	87.0	88.5	87.0	83.8	92.6
что представлять себя / what is	0.30	58.5	46.3	51.8	52.3	58.5
что происходить с / what does happen to	0.28	64.6	58.6	78.2	64.1	86.8
с какой год / starting from which year	0.28	93.9	93.9	93.9	93.9	93.9
в какой век / in which century	0.26	87.7	89.2	90.8	86.9	90.8
в какой период / in which period	0.26	86.8	83.9	88.1	82.1	86.8
к что приводить / what does X lead to	0.24	83.6	72.0	75.9	79.7	70.8
от что зависеть / what does X depend on	0.20	78.8	73.6	79.2	84.0	92.5
в какой страна / in which country	0.18	97.8	97.8	91.3	94.4	100.0
Test set		77.8	72.2	79.5	75.0	84.8

Table 12: Model F1 scores depending on questions’ leading trigrams (trigrams are lemmatized)

- Some questions require *reasoning* and *co-reference resolution*.
- A small fraction of questions uses *synonyms* and *paraphrases* that are not directly borrowed from the paragraph.
- A relatively large fraction of ‘difficult’ questions contains *misspellings* and imply *yes/no* answers.

The categorization of the sample is summarized in Table 10. One can see from the table that most potential causes of degraded performance can be attributed to poor data quality: Only 25% of cases can be explained by a need to deal with linguistic phenomena such as co-reference resolution, reasoning, and paraphrase detection.

6. Conclusions

In this study, we conducted an in-depth analysis of the Russian reading comprehension dataset SberQuAD, which was created in 2017 but was neither properly documented nor presented to the scientific community. SberQuAD creators generally followed a procedure described by the SQuAD authors, which resulted in similarly high lexical overlap between questions and sentences with answers. Our analysis demonstrates that models perform better when such overlap is high.

Despite the similarities between datasets, all the models perform worse on SberQuAD than on SQuAD, which can be attributed to having only a single answer variant and fewer answers that are named entities. Furthermore, SberQuAD

annotations might have been of poorer quality, but it is hard to quantify.

We believe that the provided analysis constitutes an important contribution to research in multilingual QA. It facilitates further studies by evaluating off-the-shelf models for reading comprehension task in Russian and identifying shortcomings related to dataset creation. The latter can serve as a guidance for improving/extension of the dataset in the future.

Acknowledgements. We thank Peter Romov, Vladimir Suvorov, and Ekaterina Artemova (Chernyak) for providing us with details about SberQuAD preparation. We also thank Natasha Murashkina for initial data processing.

7. References

- Artetxe, M., Ruder, S., and Yogatama, D. (2019). On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.
- Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., and Wang, T. (2016). MS MARCO: A Human Generated MACHine Reading COMprehension Dataset. *arXiv preprint arXiv:1611.09268*.
- Burtsev, M., Seliverstov, A., Airapetyan, R., Arkhipov, M., Baymurzina, D., Bushkov, N., Gureenkova, O., Khakhulin, T., Kuratov, Y., Kuznetsov, D., et al. (2018). Deepapavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127.

- Chen, D., Bolton, J., and Manning, C. D. (2016). A thorough examination of the CNN/Daily Mail reading comprehension task. *arXiv preprint arXiv:1606.02858*.
- Chen, D., Fisch, A., Weston, J., and Bordes, A. (2017). Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.
- Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., and Zettlemoyer, L. (2018). QuAC: Question Answering in Context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Clark, C. and Gardner, M. (2017). Simple and effective multi-paragraph reading comprehension. *arXiv preprint arXiv:1710.10723*.
- Dang, H. T., Kelly, D., and Lin, J. J. (2007). Overview of the TREC 2007 Question Answering Track. In *Proceedings of The Sixteenth Text REtrieval Conference (TREC 2007)*.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Ferrucci, D. A., Brown, E. W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J. M., Schlaef, N., and Welty, C. A. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3):59–79.
- Giampiccolo, D., Forner, P., Herrera, J., Peñas, A., Ayache, C., Forascu, C., Jijkoun, V., Osenova, P., Rocha, P., Sacaleanu, B., and Sutcliffe, R. (2008). Overview of the CLEF 2007 Multilingual Question Answering Track. In *Advances in Multilingual and Multimodal Information Retrieval*, pages 200–236.
- Hardalov, M., Koychev, I., and Nakov, P. (2019). Beyond English-Only Reading Comprehension: Experiments in Zero-Shot Multilingual Transfer for Bulgarian. *arXiv preprint arXiv:1908.01519*.
- He, W., Liu, K., Liu, J., Lyu, Y., Zhao, S., Xiao, X., Liu, Y., Wang, Y., Wu, H., She, Q., et al. (2017). DuReader: a Chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073*.
- Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.
- Hirschman, L. and Gaizauskas, R. J. (2001). Natural language question answering: the view from here. *Natural Language Engineering*, 7(4):275–300.
- Jia, R. and Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.
- Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. (2017). TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.
- Kuratov, Y. and Arhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.
- Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Li, P., Li, W., He, Z., Wang, X., Cao, Y., Zhou, J., and Xu, W. (2016). Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. *arXiv preprint arXiv:1607.06275*.
- Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4996–5001.
- Prager, J. M. (2006). Open-domain question-answering. *Foundations and Trends in Information Retrieval*, 1(2):91–231.
- Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.
- Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don’t know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.
- Reddy, S., Chen, D., and Manning, C. D. (2019). CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.
- Rondeau, M.-A. and Hazen, T. J. (2018). Systematic Error Analysis of the Stanford Question Answering Dataset. In *Proceedings of the Workshop on Machine Reading for Question Answering*.
- Seo, M., Kembhavi, A., Farhadi, A., and Hajishirzi, H. (2016). Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.
- Soboleva, D. and Vorontsov, K. (2019). Three-stage question answering system with sentence ranking.
- Sugawara, S., Kido, Y., Yokono, H., and Aizawa, A. (2017). Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Sun, Y. and Xia, T. (2019). A hybrid network model for Tibetan question answering. *IEEE Access*, 7:52769–52777.
- Talmor, A. and Berant, J. (2019). MultiQA: An Empirical Investigation of Generalization and Transfer in Reading Comprehension. *arXiv preprint arXiv:1905.13453*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.
- Wadhwa, S., Chandu, K. R., and Nyberg, E. (2018). Comparative analysis of neural qa models on squad. *arXiv preprint arXiv:1806.06972*.
- Wang, W., Yang, N., Wei, F., Chang, B., and Zhou, M. (2017). Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Compu-*

tational Linguistics (Volume 1: Long Papers), volume 1, pages 189–198.

Zhang, X., Yang, A., Li, S., and Wang, Y. (2019). Machine reading comprehension: a literature review. *CoRR*, abs/1907.01686.